Anchor regression: Heterogeneous data meet causality

Rothenhausler, D., Meinshausen, N., Bu hlmann , P., and Peters, J

Presenter: Will Zhang

University of Washington

December 12, 2022

Table of contents

Introduction

Problem: Covariates on a data set differs in distribution between training and test data (prediction). How to optimize predictive accuracy?

- Causal parameters are optimal only if test distribution varies a lot Subpar performance (too conservative) on moderately shifted data
- OLS can have high predictive error under strong intervention

Solution: Anchor regression: interpolation between OLS and 2SLS

- Robustness against linear shifts on specific sets
- Protection against intervention up to a size
- Doesn't require IV assumptions
- Improved replicability, and possible stability results

Contribution

Given centered variables $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $A \in \mathbb{R}^q$, and P_A denote the L_2-projection. For $\gamma > 0$, define population anchor regression

$$b^\gamma := \underset{b}{\operatorname{argmin}}\ \mathbb{E}_{\text{train}}\left[\left((\mathrm{Id} - \mathrm{P}_A)\left(Y - X^\top b\right)\right)^2\right] + \gamma \mathbb{E}_{\text{train}}\left[\left(\mathrm{P}_A\left(Y - X^\top b\right)\right)^2\right] \tag{1.1}$$

Define matrix containing observations: $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{A} \in \mathbb{R}^{n\times q}$ and $\mathbf{Y} \in \mathbb{R}^n$. We have plug-in estimator

$$\hat{b}^\gamma = \underset{b}{\operatorname{argmin}} \left\|(\mathrm{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\right\|_2^2 + \gamma \left\|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\right\|_2^2 \tag{1.2}$$

where $\Pi_{\mathbf{A}} \in \mathbb{R}^{n\times n}$ is projection matrix. i.e. $\Pi_{\mathbf{A}} := \mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$.

Concepts utilized

$$
\begin{aligned}
b_{\mathrm{PA}} &:= \underset{b}{\operatorname{argmin}} \mathbb{E}_{\text{train}} \left[\left((\mathrm{Id} - \mathrm{P}_A) \left(Y - X^\top b \right) \right)^2 \right] \\
&= \underset{b}{\operatorname{argmin}} \mathbb{E}_{\text{train}} \left[\left((Y - \mathrm{P}_A Y) - (X - \mathrm{P}_A X)^\top b \right)^2 \right] \\
b_{\mathrm{OLS}} &:= \underset{b}{\operatorname{argmin}} \mathbb{E}_{\text{train}} \left[\left(Y - X^\top b \right)^2 \right] \\
b_{\mathrm{IV}} &:= \underset{b}{\operatorname{argmin}} \mathbb{E}_{\text{train}} \left[\left(\mathrm{P}_A \left(Y - X^\top b \right) \right)^2 \right]
\end{aligned} \tag{1.3}
$$

$$
\begin{aligned}
b^0 &= b_{\mathrm{PA}} \\
b^1 &= b_{\mathrm{OLS}} \\
b^{\to\infty} := \lim_{\gamma \to \infty} b^\gamma &= b_{\mathrm{IV}}
\end{aligned} \tag{1.4}
$$

Linear structural causal model

Let the distribution of (X, Y, H, A) under $\mathbb{P}_{\text{train}}$ be a solution of the SEM

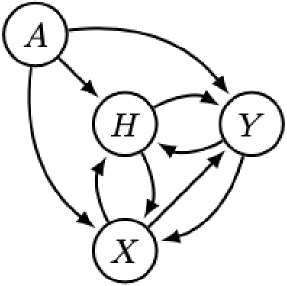
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{M}A \tag{1.5}$$

where $H \in \mathbb{R}^r$ is hidden variable, $\varepsilon \in \mathbb{R}^{d+1+r}$ is noise, $\mathbf{M} \in \mathbb{R}^{(d+1+r)\times q}$ and $\mathbf{B} \in \mathbb{R}^{(d+1+r)\times(d+1+r)}$ are unknown constant matrices.

Assuming $\mathrm{Id} - \mathbf{B}$ is invertible. Then distribution of (X, Y, H, A) is well-defined in terms of $\mathbf{B}, \varepsilon, \mathbf{M}$ and A as equation (1.5) has unique solution

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (\mathrm{Id} - \mathbf{B})^{-1}(\varepsilon + \mathbf{M}A) \tag{1.6}$$

Directed graph G



- Allow G being cyclic
- Acyclic G implies $\mathrm{Id} - \mathbf{B}$ is always invertible
- A is not an instrument
- Predictive guarantees apply to X, Y, H

Shift intervention

The new interventional distribution is denoted by $\mathbb{P}_v$. The distribution of the variables (X, Y, H) under $\mathbb{P}_v$ is defined as the solution of

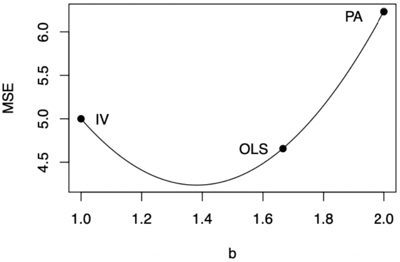
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + v \tag{1.7}$$

where the shift $v \in \mathbb{R}^{d+1+q}$ is a random or deterministic vector independent of ε, and has form $\mathbf{M}\delta$ for some δ.

Anchor regression: an example

$A \sim$ Rademacher

$\mathbb{P}_{\text{train}}$	$\mathbb{P}_v$ with $v = (1.8, 0, 0)$
$\varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{\text{indep.}}{\sim} \mathcal{N}(0,1)$	$\varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{\text{indep.}}{\sim} \mathcal{N}(0,1)$
$H \leftarrow \varepsilon_H$	$H \leftarrow \varepsilon_H$
$X \leftarrow A + H + \varepsilon_X$	$X \leftarrow 1.8 + H + \varepsilon_X$
$Y \leftarrow X + 2H + \varepsilon_Y$	$Y \leftarrow X + 2H + \varepsilon_Y$



Example continued: Performance trade off

We want to avoid overfitting. Consider following minimax loss:

$$\operatorname*{argmin}_{b} \sup_{v \in C} \mathbb{E}_v \left[\left(Y - X^\top b \right)^2 \right] \text{ for a suitable set } C \subseteq \mathbb{R}^{d+q+1}. \tag{1.8}$$

- b_{PA} solves the problem for $C_{\mathrm{PA}} = \{0\}$
- b_{OLS} solves the problem for $C_{\mathrm{OLS}} = \left\{ v \in \mathbb{R}^3 : v_2 = v_3 = 0 \text{ and } v_1^2 \leq \mathbb{E}_{\text{train}} \left[A^2 \right] \right\}$
- b_{IV} solves the problem for $C_{\mathrm{IV}} = \left\{ v \in \mathbb{R}^3 : v_2 = v_3 = 0 \right\}$

Example continued: Proof

$A \sim$ Rademacher $\implies Var(A) = 1, \mathbb{E}[A] = 0, \mathbb{E}[AA^\top] = 1.$
For $v = (v_1, v_2, v_3)$, where $v_2 = v_3 = 0$, so there is only perturbation in X.
From the setup of the example: $\mathbf{M} = (1, 0, 0)^\top$; Therefore,

$$\mathbf{M}\mathbb{E}_{\text{train}}\left[AA^\top\right]\mathbf{M}^\top = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Taking Steps from proof of Theorem 1 and Theorem 7,

$$\sup_{v \in C^\gamma} \mathbb{E}_v\left[\left(Y - X^\top b\right)^2\right] = \mathbb{E}_0\left[\left(Y - X^\top b\right)^2\right] + \sup_{v \in C^\gamma} w^\top \mathbb{E}_v\left[vv^\top\right] w$$

where $w = \left((\mathrm{Id} - \mathbf{B})^{-1}_{d+1,\bullet} - b^\top (\mathrm{Id} - \mathbf{B})^{-1}_{1:d,\bullet}\right)^\top$ is a parameter.

Example continued: Proof

From previous page, and by definition of C^γ, we know $\sup_{v\in C^\gamma} \mathbb{E}_v\left[vv^\top\right] = \gamma \mathbf{M}\mathbb{E}_{\text{train}}\left[AA^\top\right]\mathbf{M}^\top$ solves the criterion function. Hence, in this deterministic case:

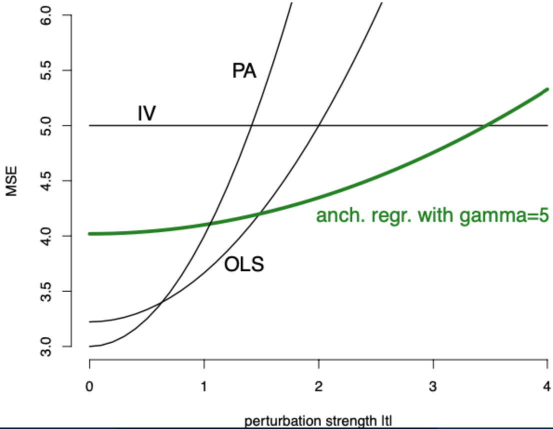
$$\sup_{v\in C^\gamma} vv^\top = \begin{pmatrix} \gamma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- $C_{\mathrm{PA}} = \{0\} \implies v \leq 0$, so only $\gamma = 0$ satisfy the definition.
- Given C_{OLS}, $v_1^2 \leq \mathbb{E}_{train}[A^2] = 1 = \gamma$
- Given C_{IV}, v_1 is arbitrary, so only $\gamma \to \infty$ satisfy the definition.

And results follow. $\square$

Example continued: Optimal performance

$\mathbb{E}_v\left[\left(Y - X^\top b\right)^2\right]$ **is depicted under perturbation** $v = (t, 0, 0)^\top$**.**



Theorem 1

Let the assumptions of (1.7) hold. For any $b \in \mathbb{R}^d$ we have

$$\mathbb{E}_{\text{train}}\left[\left((\mathrm{Id}-\mathrm{P}_A)\left(Y-X^\top b\right)\right)^2\right]+\gamma\mathbb{E}_{\text{train}}\left[\left(\mathrm{P}_A\left(Y-X^\top b\right)\right)^2\right]=\sup_{v\in C^\gamma}\mathbb{E}_v\left[\left(Y-X^\top b\right)^2\right] \quad (1.9)$$

where $C^\gamma := \left\{v \in \mathbb{R}^{d+q+1} \text{ such that } vv^\top \preceq \gamma\mathbf{M}\mathbb{E}_{\text{train}}\left[AA^\top\right]\mathbf{M}^\top\right\}$

- The squared L_2-risk under certain worst-case shift interventions is equal to adding a penalty to the risk.
- As population anchor regression optimizes the penalized criterion, anchor regression minimizes the worst-case MSE under shift interventions up to a given strength in certain directions

Theorem 7

For any $b \in \mathbb{R}^d$ we have

$$\mathbb{E}_{\text{train}}\left[\left((\mathrm{Id}-\mathrm{P}_A)\left(Y-X^\top b\right)\right)^2\right]+\gamma\mathbb{E}_{\text{train}}\left[\left(\mathrm{P}_A\left(Y-X^\top b\right)\right)^2\right]=\sup_{\mathbb{P}_v\in C^\gamma}\mathbb{E}_v\left[\left(Y-X^\top b\right)^2\right]$$

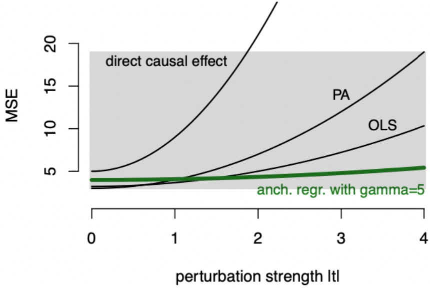
where

$C^\gamma := \{$ probability measures $\mathbb{P}_v$: the assumptions of Section 2.1 are satisfied, and $\mathbb{E}_v\left[vv^\top\right] \preceq \gamma \mathbf{M}\mathbf{E}_{\text{train}}\left[AA^\top\right]\mathbf{M}^\top\}$.

Limitations of direct causal effect

For the perturbed distribution $\mathbb{P}_v$ is given under a shift $v = (0,0,t)^\top, t \in \mathbb{R}$. $A \sim$ Rademacher

$$\begin{array}{cc} \varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{\text{indep.}}{\sim} \mathcal{N}(0,1) & \varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{\text{indep.}}{\sim} \mathcal{N}(0,1) \\ H \leftarrow A + \varepsilon_H & H \leftarrow t + \varepsilon_H \\ X \leftarrow H + \varepsilon_X & X \leftarrow H + \varepsilon_X \\ Y \leftarrow 1 \cdot X + 2H + \varepsilon_Y & Y \leftarrow 1 \cdot X + 2H + \varepsilon_Y \end{array}$$



Interpretation of anchor regression via quantiles

Define $Q(\alpha)$ as the α-th quantile of $\mathbb{E}\left[\left(Y - X^\top b\right)^2 \mid A\right]$.

Lemma 1. Assume that the variables (X, Y, A) follow a centered multivariate normal distribution under $\mathbb{P}$. Then, for $0 \leq \alpha \leq 1$,

$$Q(\alpha) = \mathbb{E}\left[\left((\mathrm{Id} - \mathrm{P}_A)\left(Y - X^\top b\right)\right)^2\right] + \gamma \mathbb{E}\left[\left(\mathrm{P}_A\left(Y - X^\top b\right)\right)^2\right] \tag{1.10}$$

where γ equals the α-th quantile of a χ^2-distributed random variable with one degree of freedom.

Projectability condition & lemma 2

Say the projectability condition is fullfilled if

$$\operatorname{rank}\left(\mathrm{Cov}_{\mathrm{train}}\left(A, X\right)\right) = \operatorname{rank}\left(\mathrm{Cov}_{\mathrm{train}}\left(A, X\right) \mid \mathrm{Cov}_{\mathrm{train}}\left(A, Y\right)\right), \tag{2.1}$$

where $\mathrm{Cov}_{\mathrm{train}}\left(A, X\right) \mid \mathrm{Cov}_{\mathrm{train}}\left(A, Y\right)$ is a $q \times (d+1)$ matrix.

- Projectability condition generally allows that the anchor variables A directly influence also Y or H
- EX: $\mathrm{Cov}_{\mathrm{train}}(A, X)$ is of full rank and $q \leq d$
- If $q > d$, additional constraints on $A \to Y$ is required

Lemma 2. Assume that $\mathbb{E}_{\mathrm{train}}\left[AA^\top\right]$ is invertible. The projectability condition (2.1) is fulfilled if and only if

$$\min_b \mathbb{E}_{\mathrm{train}}\left[\left(\mathrm{P}_A\left(Y - X^\top b\right)\right)^2\right] = 0. \tag{2.2}$$

Replicability of $b^{\to\infty}$

Consider two different data-generating distributions, denoted by "train" and "test" (with prime on variables)

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + v, v = \mathbf{M}\delta, \delta = \kappa A + \xi, \tag{2.3}$$

where ξ is a random vector with mean zero and independent of ε and A and $\kappa \neq 0$.

- v' and A' can have arbitrarily different distributions than v and A
- **B** and **M** are the same

$$\mathrm{Cov}_{\text{test}}\left(\varepsilon'\right) = L\,\mathrm{Cov}_{\text{train}}\left(\varepsilon\right) \text{ for } L > 0, \mathbb{E}_{\text{test}}\left[\varepsilon'\right] = \mathbb{E}_{\text{train}}\left[\varepsilon\right] = 0 \tag{2.4}$$

Replicability of $b^{\to\infty}$

Consider the parameter $b^{\to\infty}$ as defined in (1.4),

$$
\begin{aligned}
b^{\to\infty} &= \underset{b \in I}{\operatorname{argmin}} \mathbb{E}_{\text{train}} \left[\left(Y - X^\top b \right)^2 \right], \\
I &= \left\{ b; \mathbb{E}_{\text{train}} \left[Y - X^\top b \mid A \right] \equiv 0 \right\},
\end{aligned}
$$

similar for $b'^{\to\infty}$

Theorem 2 Consider the models in (2.3), for the training and test data, respectively. Assume (2.4) and $\mathbb{E}_{\text{train}} \left[AA^\top \right]$ and $\mathbb{E}_{\text{test}} \left[A' \left(A' \right)^\top \right]$ are invertible and assume that the projectability condition (2.1) holds. Then,

$$
b'^{\to\infty} = b^{\to\infty}.
$$

Anchor stability

Anchor stable: all solutions of anchor regression agree (i.e., if $b^0 = b^\gamma$ for all $\gamma \in [0, \infty)$)

- Predictive stability and replicability of variable selection under certain perturbations
- Allows a causal interpretation of the coefficient vector under otherwise comparatively weak assumptions

Proposition 1. If $b^0 = b^{\to\infty}$ then

$$b^0 = b^\gamma \text{ for all } \gamma \in (0, \infty).$$

Anchor stability: case

Given $b^0 = b^{\to\infty}$, we have

$$\mathbb{E}_{\text{train}}\left[\left((\mathrm{Id}-\mathrm{P}_A)\left(Y - X^\top b\right)\right)^2\right] = \mathbb{E}_{\text{train}}\left[\left(\mathrm{P}_A\left(Y - X^\top b\right)\right)^2\right]$$
$$\mathbb{E}_{\text{train}}\left[\left(Y - X^\top b\right)^2\right] = 2\mathbb{E}_{\text{train}}\left[\left(\mathrm{P}_A\left(Y - X^\top b\right)\right)^2\right]$$

- Independence of A, Multivariate Normal, etc. do not satisfy the condition.
- zero variance case works but unrealistic...

Anchor stability

Theorem 3 Let the assumptions in Linear structural causal model hold, and in addition assume the projectability condition (2.1) and that the Gram matrix $\mathbb{E}_{\text{train}}\left[AA^\top\right]$ is invertible. If $b^0 = b^{\to\infty}$, then, for all random or constant vectors v that are uncorrelated of ε and take values in span($\mathbf{M}$),

1. $\mathbb{E}_{\text{train}}\left[\left(Y - X^\top b^0\right)^2\right] = \mathbb{E}_v\left[\left(Y - X^\top b^0\right)^2\right]$, and
2. $b^0 = \text{argmin}_b\, \mathbb{E}_v\left[\left(Y - X^\top b\right)^2\right]$.

Anchor stability

Theorem 4 Let the assumptions in Linear structural causal model hold with an acyclic graph G, and assume the projectability condition (2.1).

Furthermore, assume that for every disjoint sets of variables $V_1, V_2, V_3 \subset (X, Y, H, A)$, V_1 is d-separated of V_2 in G given V_3 if and only if the partial correlation $\text{part.cor}\left(V_1, V_2 \mid V_3\right) = 0$. Furthermore assume that for each X_k there exists k' such that $A_{k'} \rightarrow X_k$. If $b^{\rightarrow\infty} = b^0$, then

$$b^{\rightarrow\infty} = b^0 = \partial_x \mathbb{E}[Y \mid do(X = x)]$$

In addition, there is no confounder between X and Y, i.e., there is no H_k that is both an ancestor of some $X_{k'}$ and Y in G

- Anchor stability has causal interpretation & no confounder
- A positive indication for replicability

Estimator in the low-dimensional setting

Concatenating the observations row-wise forms matrices that we denote $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{Y}$. Assume b^γ is unique. For $d < n$, use a plug-in estimator:

$$\hat{b}^\gamma = \underset{b}{\operatorname{argmin}} \left\| (\operatorname{Id} - \Pi_{\mathbf{A}}) (\mathbf{Y} - \mathbf{X}b) \right\|_2^2 + \gamma \left\| \Pi_{\mathbf{A}} (\mathbf{Y} - \mathbf{X}b) \right\|_2^2, \tag{3.1}$$

where $\Pi_{\mathbf{A}} \in \mathbb{R}^{n \times n}$ is the matrix that projects on the column space of $\mathbf{A}$, i.e., $\Pi_{\mathbf{A}} := \mathbf{A} \left(\mathbf{A}^\top \mathbf{A} \right)^{-1} \mathbf{A}^\top$. To write in OLS form, define

$$\tilde{\mathbf{X}} := (\operatorname{Id} - \Pi_{\mathbf{A}}) \mathbf{X} + \sqrt{\gamma} \Pi_{\mathbf{A}} \mathbf{X} \quad \text{and} \quad \tilde{\mathbf{Y}} := (\operatorname{Id} - \Pi_{\mathbf{A}}) \mathbf{Y} + \sqrt{\gamma} \Pi_{\mathbf{A}} \mathbf{Y}. \tag{3.2}$$

The estimator in (3.1) can be represented as follows:

$$\hat{b}^\gamma = \underset{b}{\operatorname{argmin}} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} b\|_2^2.$$

Distributional results for $\hat{b}^\gamma - b^\gamma$ are not investigated.

Estimator in the high-dimensional setting

If $d \gg n$, the plug-in estimator is not well defined. Assume $p < n$ s.t. Π_A is well-posed. Propose

$$\hat{b}^{\gamma,\lambda} = \underset{b}{\operatorname{argmin}} \, \|(\mathrm{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2 + \gamma \, \|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2 + 2\lambda\|b\|_1. \tag{3.3}$$

which induces sparsity by ℓ_p-norm and parameter λ.
With transformation in (3.2), regularized anchor regression can be rewritten

$$\underset{b}{\operatorname{argmin}} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}b\|_2^2 + 2\lambda\|b\|_1,$$

Finite-sample bound for discrete anchors

For all $A = a, a \in \mathcal{A}$ are given equal weight. Objective function becomes

$$R(b) := \mathbb{E}_{\text{train}} \left[\left(Y - X^\top b - \mathbb{E}_{\text{train}} \left[Y - X^\top b \mid A \right]\right)^2 \right] + \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left(\mathbb{E}_{\text{train}} \left[Y - X^\top b \mid A = a \right] \right)^2$$

Denote n_a for the number of observations in $A = a$ and $n_{\min} := \min_{a \in \mathcal{A}} n_a$. We write $\mathbf{X}^{(a)} \in \mathbb{R}^{n_a \times d}$ consisting of row observations $\mathbf{X}_{i,\bullet}$ for $\mathbf{A}_i = a$. $\overline{\mathbf{X}}^{(a)} = \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbf{X}_{i,\bullet}^{(a)}$. Analogously we define $\mathbf{Y}^{(a)} \in \mathbb{R}^{n_a}$ and $\overline{\mathbf{Y}}^{(a)}$. High-dimensional anchor regression estimator in (3.3) with equal weight equals

$$\hat{b} := \underset{b}{\operatorname{argmin}} \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} \left(\mathbf{Y}_i^{(a)} - \overline{\mathbf{Y}}^{(a)} - \left(\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)} \right) b \right)^2 + \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left(\overline{\mathbf{Y}}^{(a)} - \overline{\mathbf{X}}^{(a)} b \right)^2 + 2\lambda \|b\|_1$$

Anchor compatibility constant

For any $S \subseteq \{1, \ldots, d\}$ and stretch factor $L > 0$ define the anchor compatibility constant

$\hat{\phi}^2(L, S) :=$

$$\min_{\|b_S\|_1 = 1, \|b_{-S}\|_1 \le L} |S| \Big(\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} \left(\left(\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)} \right) b \right)^2 + \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left(\overline{\mathbf{X}}^{(a)} b \right)^2 \Big)$$

To proceed, we need a lower bound on the compatibility constant $\hat{\phi}^2 (L, S^*)$ for $S^* := \left\{ k : b_k^\gamma \neq 0 \right\}$, the active set of b^γ. Note that for all S

$$\hat{\phi}^2(L, S) \ge \min(\gamma, 1) \min_{\|b_S\|_1 = 1, \|b_{-S}\|_1 \le L} \frac{|S|}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} \left(\mathbf{X}_{i,\bullet}^{(a)} b \right)^2 .$$

Theorem 5

Consider the model in (8) and assume that ε is multivariate Gaussian. Moreover, assume that $\left(\mathbf{X}_{i,\bullet}^{(a)}, \mathbf{Y}_i^{(a)}\right)$, $i = 1, \ldots, n_a$, are i.i.d. random variables that follow the distribution of $(X, Y) \mid A = a$ under $\mathbb{P}_{\text{train}}$. Fix $\gamma > 0$ and assume that $\hat{\phi}^2(8, S^*) \geq c$ for some constant $c > 0$ with probability $1 - \delta$, and that $S^* \neq \emptyset$. Choose $t \geq 0$ such that

$$|S^*|^2 \left(t + \log(d) + \log(|\mathcal{A}|)\right)/n_{\min} \leq c',$$

for some constant $c' > 0$. Then, for $\lambda \geq C\sqrt{(t + \log(d) + \log(|\mathcal{A}|))/n_{\min}}$, with probability exceeding $1 - 10\exp(-t) - \delta$

$$R(\hat{b}) \leq \min_b R(b) + C'\lambda^2 |S^*|,$$

where constants $C, C' < \infty$ depend on $\max_k \left(\mathrm{Var}\left(X_k\right), \mathrm{Var}\left(Y - X^\top b^\gamma\right)\right)$, $\max_{a\in\mathcal{A}} \|\mathbb{E}_{\text{train}}\left[X \mid A = a\right]\|_\infty, \max_{a\in\mathcal{A}} \left|\mathbb{E}_{\text{train}}\left[Y - X^\top b^\gamma \mid A = a\right]\right|), \gamma, c$ and c'. The variances are meant with respect to the measure $\mathbb{P}_{\text{train}}$.

EX1: Genotype-tissue expression

Consider gene expressions from 13 different tissues, Goal: find relevant features that can be found on the other tissues

- Y =expression of target gene, X =expressions of all other genes
- $y \in \{1, ..., d\}$ are target genes' indices, $x \in \{1, ..., d\} \setminus y$ are all other genes' indices
- For each tissue, additional covariates (genotyping principal components, PEER factors, sex, and genotyping platform) as anchors (account for patch effect, environmental effect, history etc.)

1. compute and rank variables using the Lasso and penalized anchor regression on one specific tissue t
2. check whether the discoveries can also be replicated on the other tissues $t' \neq t$

EX1: Genotype-tissue expression

Ranking by anchor stability improves replicability and $b^{\gamma \to \infty}$ is unstable under weak correlation. Thus, check whether AR coefficients are bounded away from 0 for $\gamma \in [0,1]$

Consider tissue t. For AR, compute

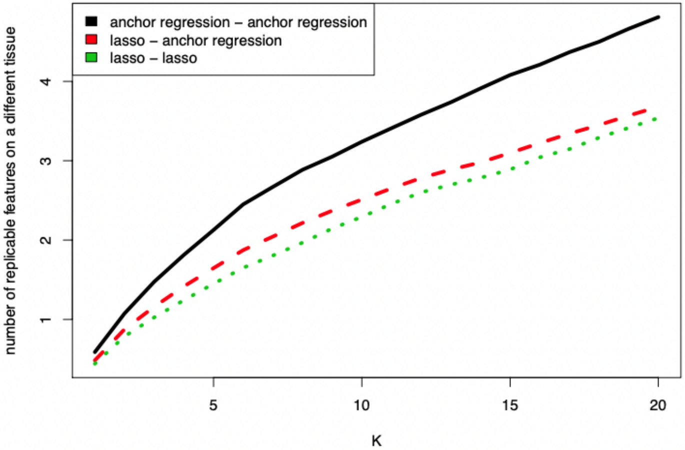
$$a_{y,k,t} := \min_{\gamma \in [0,1]} \left| \hat{b}_k^{\gamma,\lambda} \right|, \tag{4.1}$$

where $\hat{b}^{\gamma,\lambda}$ is the $p-1$-dimensional anchor coefficient of $y \in \{1,\dots,p\}$ on the other gene expressions $x = \{1,\dots,p\} \setminus \{y\}$. For comparison, compute the Lasso coefficients

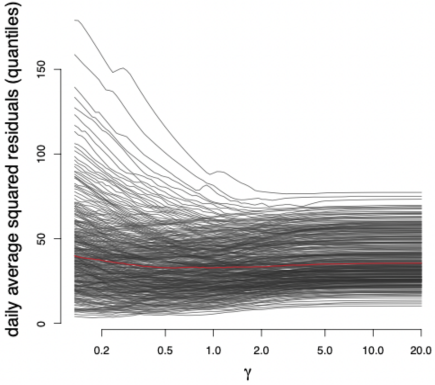
$$l_{y,k,t} := \left| \left(\hat{b}_{\text{lasso}} \right)_k \right|, \tag{4.2}$$

$\hat{b}_{\text{lasso}} = \hat{b}^{0,\lambda} \implies a_{y,k,t} \le l_{y,k,t}$.

EX1: Improved replicability with stable anchor regression



EX2: Bike sharing data

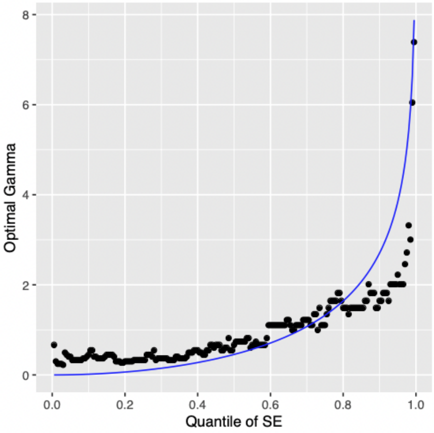


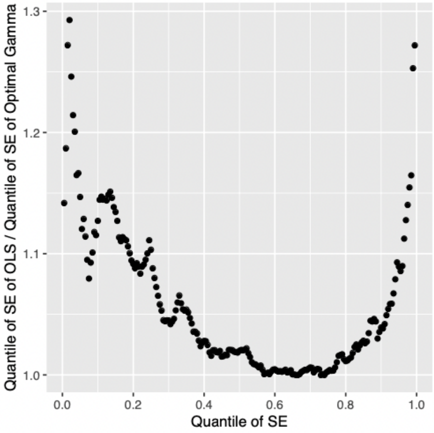
Data: $n = 17379$ hourly counts of bike rentals

Goal: predict bike rental counts using weather data reliably across days

- Use "date" as anchor (discrete with one level per day
- Use covariates temperature, feeling temperature, humidity and windspeed to predict

EX2: Optimal γ and predictive performance





Practical guidance

Possible Applications.
Generalizing across heterogeneous data, including batch effects, population shifts, and heterogeneity across time or locations.

Choice of the anchor variable.
Main prediction assumptions: linearity of the system and exogeneity of the anchor. Choose variables that we aim to achieve robustness or invariance across.

Choice of the regularization parameter.
Based on subject matter knowledge or cross-validation.

Limitations.
Assumption of linearity. C^{γ} does not contain arbitrary shifts.

Outlook

Beyond shift interventions.
Penalty schemes arises from other types of perturbations, such as noise, edge functions and do-interventions.
Nonlinear models.
Using bias-variance decomposition and assume constant variance conditional on A, we define solution

$$g^{\gamma} := \arg\min_{g \in \mathcal{G}} \mathbb{E}_{\text{train}} \left[\left((\mathrm{Id} - P_A)(Y - g(X)) \right)^2 \right] + \gamma \mathbb{E}_{\text{train}} \left[\left(P_A(Y - g(X)) \right)^2 \right]$$

Thank you

d-Separation

A path p is said to be *d*-separated (or blocked) by a set of nodes Z if and only if

1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or
2. p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set Z is said to *d*-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

More on `https://www.andrew.cmu.edu/user/scheines/tutor/d-sep.html#explanation`

Backdoor

A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if:

1. no node in Z is a descendant of X_i; and
2. Z blocks every path between X_i and X_j that contains an arrow into X_i.

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

Back-Door Adjustment

If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula

$$P(y \mid \hat{x}) = \sum_{z} P(y \mid x, z) P(z)$$